Why reliabilism is not enough

Why Reliabilism Is Not Enough: Epistemic and Moral Justification in Machine Learning

Andrew Smart, Larry James, Ben Hutchinson, Simone Wu, Shannon Vallor
Google
1600 Amphitheatre Parkway
Mountain View, California, 94043
andrewsmart@google.com

ABSTRACT

In this paper we argue that standard calls for explainability that focus on the epistemic inscrutability of black-box machine learning models may be misplaced. If we presume, for the sake of this paper, that machine learning can be a source of knowledge, then it makes sense to wonder what kind of *justification* it involves. How do we rationalize on the one hand the seeming justificatory black box with the observed widespread adoption of machine learning? We argue that, in general, people implicitly adopt *reliabilism* regarding machine learning. Reliabilism is an epistemological theory of epistemic justification according to which a belief is warranted if it has been produced by a reliable process or method [18]. We argue that, in cases where model deployments require *moral* justification, reliabilism is not sufficient, and instead justifying deployment requires establishing robust human processes as a moral "wrapper" around machine outputs. We then suggest that, in certain high-stakes domains with moral consequences, reliabilism does not provide another kind of necessary justification—moral justification. Finally, we offer cautions relevant to the (implicit or explicit) adoption of the reliabilist interpretation of machine learning.

ACM Reference Format:
Andrew Smart, Larry James, Ben Hutchinson, Simone Wu, Shannon Vallor, Google, 1600 Amphitheatre Parkway, Mountain View, California, 94043, andrewsmart@google.com . 2020. Why Reliabilism Is Not Enough:, Epistemic and Moral Justification in Machine Learning. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20), February 7–8, 2020, New York, NY, USA*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3375627.3375866

1 INTRODUCTION

Epistemology is the systematic philosophical examination of knowledge and is concerned with the nature of knowledge and how we acquire it [25]. Amongst philosophers, there is consensus that for a mental state to count as a knowledge state it must minimally be a justified, true belief. If we presume, for the sake of this paper, that machine learning can be a source of knowledge, then it makes sense to wonder what kind of justification it involves. *Prima facie*, one might think that machine learning is epistemologically inscrutable [38]. After all, we don't usually have access to the black box in which models make decisions. Thus it might appear that machine learning decisions *qua* knowledge don't have sufficient justification to count as knowledge. One might think this is because the models don't appear to have evidence or accessible reasons for their output. We suggest that this underlies the widespread interest in explainable or interpretable AI within the research community as well as the general public. Despite this inscrutability, machine learning is being deployed in human-consequential domains at a rapid pace. Reliabilism is an epistemological theory of epistemic justification according to which a belief is warranted if it has been produced by a reliable process or method [18]. In this paper, we explore what this means in the ML context. We then suggest that, in certain high-stakes domains with moral consequences, reliabilism does not provide another kind of necessary justification—moral justification.

In this paper we argue that standard calls for explainability that focus on the epistemic inscrutability of black-box machine learning models may be misplaced. We further argue that in cases where model outputs require epistemic and moral justification, there is a need to establish robust human processes as a moral "wrapper" around machine outputs. Finally, we offer a general caution, relevant if we adopt the reliabilist interpretation of ML models, to be especially sensitive to distribution drift and to continually check the fit of the model to the population it is intended to serve.

Machine learning has become a transformative technology that is now impacting almost every aspect of life in many countries [28]. Yet it is widely acknowledged that the precise mechanisms by which machine learning generates predictions are quite mysterious [12]. This is because machine learning mixes myriad data sources together into abstract mathematical objects—such as model weights—which are not human-interpretable. Hence the models are often called "black boxes" because we cannot peer into them to find out how they work.

In this paper, we explore what we can learn about machine learning if we consider it as a knowledge-generating enterprise. We use the more specially defined sense of knowledge from the fields of epistemology and philosophy of science, which is that knowledge is justified, true belief. If machine learning is not a source of knowledge, or is not at least intended to be a source of knowledge, it seems self-negating to use machine learning in the first place. As mentioned, within epistemology the fundamental conditions for a mental state (without worrying about what mental states *are*) to be counted as knowledge are that it is minimally a justified, true belief. These conditions of justification, truth, and belief while being

AIES '20, February 7–8, 2020, New York, NY, USA

ACM ISBN 978-1-4503-7110-0/20/02…$15.00
https://doi.org/10.1145/3375627.3375866

necessary for a mental state to count as knowledge are not, however, sufficient. The Gettier Problem [16] [1] suggests that something else in addition is required, however, we set aside the debate about Gettier problems here as the focus is going to be on justification.

From a theoretical perspective in the field of machine learning, it is unclear why a certain configuration of weights or the architecture of a neural network should perform any better than another pattern. Machine learning practitioners and theoreticians go through a number of trial-and-error iterations to find which arrangement of elements and parameters in a model work best for a given task. Once the performance has been optimized it is not clear why the particular model arrangement produced optimal results. Nonetheless, these weights and the non-linear functions they represent, configured in certain layered patterns in a network, can outperform humans on some tasks and games [1]. In a keynote address at NeurIPS 2017 Ali Rahimi[2] famously likened machine learning to alchemy—in other words a prescientific discipline that accidentally produces useful results. Despite this lack of theoretical understanding, machine learning is adopted readily in many domains of life that can have severe consequences for human beings such as medicine [14], security [40], criminal justice [4], and surveillance [2], among many others.

How can we reconcile this view of machine learning as alchemy with our suggestion that we consider it to be a knowledge generating enterprise? On the basis of what epistemic grounds should human society or science trust the knowledge produced by this alchemical marvel? Because, in many domains, machine learning yields successful results and where adopted it is often reliable (consider, for example, machine translation and predictive recommendations). We argue that that the adoption of machine learning implicitly reflects a general reliabilist stance toward the outputs of machine learning algorithms. Reliabilism is an approach to justification and knowledge according to which a belief is warranted if it has been produced by a reliable process or method [33]. On the basis of reliabilism, it is enough that a belief has been the product of a reliable process for it to be justified; that is, there is no further requirement that the reliability of the process or method can be independently proven or justified by the knower. We can know things as a result of a reliable process (such as our neurobiological system of visual perception) even if the precise causes of that reliability are opaque to us.

Indeed, given the theoretical and epistemological situation in machine learning, where the technology is far ahead of scientific understanding, reliabilism seems to be the *only* available epistemic approach to justification for believing the outputs of machine learning models.

Thus, despite the fact that what justifies the outputs of machine learning models is not accessible to us, justification is not of necessity imperilled so long as those outputs are produced by a reliable process. And while the possibility of reliabilist justification should give us some comfort that we have not been flippant in accepting the outputs of models whose inner workings are opaque to us, in the second part of this paper we will argue that there are contexts where a reliable process is not up to the task of providing all the kinds of justification one might need; in particular, moral justification.

2 MACHINE LEARNING AS A KNOWLEDGE GENERATING ENTERPRISE

2.1 Types of Epistemic Justification and their Relationship to Explainability

In epistemology and philosophy of science, *justification* is the property of a true belief that converts it to knowledge [33]. In other words, justification is the process of rendering a belief warranted. The knowledge equation of knowledge = justified, true belief can come apart: beliefs can be justified, though false. Justification is a tool for maximizing true beliefs, but not an infallible one.

Let's take a minute to conduct a brief overview of common kinds of justification defended in the philosophical literature. And in so doing we'll address whether in these theories the form of justification is explainable in a way that would matter for machine learning.

Foundationalism is a form of justification which appeals to a hierarchical linear theory wherein beliefs are divided into two categories: basic (which are self-justified or self-evident) and derived (which depend on the basic beliefs and whose justification is inferential) [33]. Basic beliefs are phenomenal or the result of sense data.

Coherentism in contrast denies the division between beliefs into basic and derived. According to coherentism, all beliefs are justified insofar as the system as a whole is justified. If a belief coheres with the rest of your beliefs it is justified. A problem for coherentism is that it cannot easily explain how beliefs relate to the world. Thus some priority must be given to beliefs that are not justified exclusively on the basis of their internal connections to other beliefs. This modified version of coherentism was defended by Quine and is known as the "web of beliefs" [35].

There are also *appeals to authority* defended by philosophers over the centuries [42]. We can justify our beliefs by appealing to expert opinion or authorities on a topic. We might believe in black holes because we read a book by Stephen Hawking. But again, the problem of fallibilism comes in because even Stephen Hawking can be wrong.

These forms of justification are all explainable; we can cite the sources or reasons that render our beliefs warranted even if these forms of justification do not guarantee that our beliefs are knowledge.

However, as hinted at, within epistemology we can have justification without explainability and this is *reliabilism*. For a reliable process to be justified, there is no further requirement that the reliability of the process or method is independently proven or justified. Human perceptual processes are putative examples of a reliable process, e.g., "I know that the door is closed" because my visual system produces this perception and whenever I have this perception and I get to where the door is, it is closed. Here too the

[1] Gettier attempted to show that belief might be justified and true but not count as knowledge. This is because it is possible for justification for a belief, while counting as a legitimate source of justification, not to be justification for the specific belief in question. That is, for a particular true belief, the justification one has might seem to justify the belief but in fact it doesn't. E.g., I believe my cousin is in town because I saw him walking about. Turns out he is in town but the person I saw was his twin. So I am justified in thinking he is in town and it is true he is in town, but I don't know it to be the case.

[2] https://www.youtube.com/watch?v=x7psGHgatGM

possibility of justified false belief occurs. Psychological phenomena such as change blindness and choice blindness are common examples of justified false belief [20]. In the animal kingdom for example, because dragonflies use polarized light to guide them to water, they often land in pools of oil because oil polarizes light to a far greater degree than water [22]. Thus, dragonflies mistakenly believe pools of oil are pools of water, land in them and become stuck. For the purposes of this paper, we do not need to address debates about virtue reliabilism, internalist versus externalist interpretations of reliabilism and the many other permutations of reliabilism. [17]. However, future work on reliabilism and machine learning should address these issues.

The fundamental point about adopting a reliabilist interpretation of machine learning is this: the fact that we can't explain what's going on inside a machine learning model does not matter because reliabilism shows us that the inscrutability of a process is not a bar to its conferring justification. Conceived as embodying a reliable process, machine learning is a truth-conducive or truth-getting knowledge-generating process.[3] To an extent, this vindicates the common reliance on machine learning.[4]

3 THE RELIABILIST INTERPRETATION OF MACHINE LEARNING

Reliabilism in fact shifts the focus of epistemology from the subject's own cognition, which is taken to be transparent in terms of being able to provide reasons for beliefs, to the natural (in the case of machine learning, artificial) processes and methods by which beliefs can be gained and sustained [33]. A stronger version of reliabilism is known as *process reliabilism*. The idea of process reliabilism for theories of epistemic justification is: a belief is justified if and only if it is produced by a process that reliably leads to true belief [11].

There are increasing calls for "explainability" for machine learning algorithms because their nested non-linear structure makes the explanation for their outputs inscrutable [19, 21, 37]. This includes nascent regulatory and legislative approaches outlining a right to an explanation [36]. Researchers and regulators are shifting their focus to techniques and incentives to produce machine-learning systems that can explain themselves to their human users [30]. As Kluttz et al. argue and, on the reliabilist approach, these calls for explainability may be misplaced, at least when focusing on the epistemic justification problem rather than the moral justification problem as we will argue below.

The lack of information about what exactly makes models arrive at their predictions is unsettling, especially for domains like medicine. As humans, we search for causes and look for explanations in order to *understand* how and why things around us are the way they are, or behave in certain ways [34]. An explanation is an answer to a why question. The reliabilist approach to justification is, however, precisely justification without explainability. We are justified in accepting the outputs of a model because of the truth-indicating properties of machine learning models. On the reliabilist approach we do not need to understand the black box model to believe its output, and to have that output count as knowledge.

There are many examples used to explicate reliabilism—here is a classic one. Feldman [15] presents the following case that is pertinent to machine learning epistemology: two bird-watchers, a novice and an expert, are together in the woods when a pink-spotted flycatcher lands on branch. Both the novice and the expert believe that the bird is a pink-spotted flycatcher. Both are correct, but reliabilism argues that only the expert is justified in their belief about the bird's identity. This is because the novice is only guessing, whereas the expert is matching their vast store of knowledge in memory about bird species to their current visual experience of the bird sitting nearby. The latter process is reliable, whereas the guesses of a novice are not. Crucially, the bird-watching expert would not be able to precisely articulate the cognitive and neural processes involved in their positive identification of the bird—in other words the way in which the bird-watching expert is justified is not completely explainable. They might say something like, "I'm a bird-watching expert, that is how I know it's a pink-spotted flycatcher."

We can compare the difference between the novice and expert bird-watcher to a machine learning algorithm trained to classify birds. A trained bird-classification algorithm would have processed millions of pictures of birds, including many examples of the pink-spotted flycatcher. So even if we cannot see precisely which input features and which correlations between these features and the output caused the algorithm to identify the pink-spotted flycatcher, we know that a well-trained neural network will identify the correct bird a very high percentage of the time (because it is based on in-domain data)—thus the network tends to track truth in this case. Using a machine learning algorithm to identify a bird is reliable, and in some cases might even be more reliable than an expert bird-watcher. We can't explain exactly what the deep learning network is doing, but it does not matter in this case because the process is reliable.

Reliabilism has traditionally focused on the natural processes by which knowledge can be acquired and maintained, and therefore a diagnosis of reliabilism with regards to machine learning might be met with some skepticism. Next we consider some possible objections to reliabilism.

3.1 Objections to Using Reliabilism to Account for the Black Box

An objection is that even reliable knowledge-forming processes remain fallible—yet according to the knowledge equation (where knowledge = justified, true belief) it seems as though knowledge must be by definition *infallible* [25]. If we say S knows that P, and yet grant that there is a certain possibility that not-P it seems as if S does not after all know that P [25]. This apparent incoherence of the notion of fallible knowledge can be resolved, however, by granting that we never have fallible (actual) knowledge, yet individual knowledge *claims* may sometimes be fallible, even when they arise from generally reliable processes. This is true both for claims arising from reliable natural knowledge-generating processes such as ordinary perception, and for claims arising from reliable *artificial* knowledge-generating processes, such as reliable machine learning

[3] If machine learning is not a truth-conducive knowledge-generating process this represents a much deeper problem that is outside the scope of this paper.
[4] There are still other internalist and externalist theories of justification that we have not discussed. However, for our present purposes of opening up new lines of inquiry, it is not necessary that we be comprehensive in this survey.

models. Indeed, the stochastic nature of many machine learning models entails that model fallibility coexists with model reliability.

Naturally, we do not want to use machine learning to generate falsehoods, and we do not want to mistakenly believe falsehoods generated by machine learning models. As Leplin [24] argues, epistemic justification is justification that advances the epistemic goal of believing truths without believing falsehoods. However, it is widely known that machine learning algorithms are fragile when tested on data that was not part of the training distribution [29], and thus can be prone to generating falsehoods when deployed in real world settings. A deep learning model with billions of parameters trained and evaluated on ImageNet can be expected to be accurate on images like those in ImageNet, but for real world scenes these models become very brittle [29]. Thus we have to restrict the domains in which models are reliable.

Machine learning algorithms, ostensibly in contrast to humans, lack reasons in the strict sense for their output. Furthermore, unless a human manually retrains a new model, machine learning algorithms cannot update their predictions to include extenuating circumstances or unanticipated but relevant information. While this might make machine learning seem irrational, in the sense of being non-responsive to reasons that are salient, we argue that under the process reliabilist interpretation we can nevertheless be justified in believing the output of an algorithm.

4 ARE CALLS FOR EXPLAINABILITY MISPLACED? EPISTEMIC VERSUS MORAL JUSTIFICATION

Given the reliabilist interpretation of justification in the ML context, one might be inclined to dismiss calls for explainabiity as hanging on a naive understanding of what kinds of justification are possible. But is that right—should we be so quick to abandon the intuition that explainability matters? Several authors have argued for similar shifts away from explanations of models or algorithms themselves, to explanations of the processes and design decisions that lead to the creation, deployment and use of the algorithm [6, 30, 38].

Many machine learning models carry out functions that go beyond simple classification or entertainment. Machine learning models are now used to make many very morally consequential decisions. The much-discussed COMPAS algorithm gives prisoners a risk score which should predict the likelihood that they will recommit a crime once released (or more accurately, the likelihood they will be re-arrested [13]). This score is then used as part of the decision about whether to release the prisoner [9]. Credit scoring companies are increasingly using machine learning to determine credit risk and loan worthiness [23, 27]. Humans in positions of power then use the outputs of these models to make decisions about other humans that have severe consequences for the individual in question. They may have to stay in jail longer, they may be rejected for a loan that would have drastically altered the course of their life.

In many of the most common commercial domains of machine learning such as predicting churn among app users, content recommendation,[5] or ad targeting the moral consequences of the model's output may *prima facie* be considered less important. These domains may not require the same kind of moral justification as decisions that directly impact the life course of an individual, or in cases where models exhibit algorithmic discrimination, and reinforce a legacy of structural racism [3]. Digging deeper though, it might be possible to infer sexual orientation based on content recommendations, which may be morally consequential in contexts where some orientations are persecuted. Thus even seemingly morally inconsequential models can be morally consequential in certain contexts [39].

Even if we are epistemically justified under reliabilism in believing the outputs of machine learning without explainability, in the next section we argue that reliabilism does not provide sufficient justification in cases where moral knowledge is at stake.

4.0.1 Moral Knowledge and Moral Epistemology. Continuing along on our assumption that machine learning is a knowledge-generating enterprise, some models will be considered to generate morally relevant knowledge. Here we are thinking about those models whose outputs are human consequential (e.g., COMPAS). We do not claim that machine learning models can generate moral knowledge itself, as it is unclear that these systems can reliably track moral properties or salience. However, morally relevant outputs of machine learning models may be incorporated into moral knowledge. From an epistemological point of view, moral knowledge is just a subset of knowledge. One can be said to have moral knowledge when one's moral beliefs are true and held justifiably. We are not, however, going to wade into the debate on whether moral facts exist, even though there might be empirically compelling reasons to think so (from anthropological evidence it seems that there are core moral values shared globally) [8]. We'll avoid this because much of the value in what we are going to say only requires that it's plausible that the output of a model could contribute to moral knowledge [6].

5 REQUIREMENTS OF MORAL JUSTIFICATION

Inasmuch as we are considering the output of a machine learning model as contributing to moral knowledge, we have to consider the kind of justification that could provide warrant for such knowledge. Can there, for example, be reliabilist moral justification?

It will be instructive to look at domains where human-consequential decisions are frequently made for some inspiration. When a human decision-maker is required to make a consequential decision about other humans, there are often elaborate institutions and practices to ensure legally and socially sanctioned accountability or oversight around the decision [41]. For example, in the legal, medical and banking industries there are regulations, institutions and laws which govern the way in which decisions can be justified and grant people, at least in principle, the right to redress impactful decisions. In other words, moral decisions must be defeasible or contestable. Additionally, notions of accountability require that a decision maker must be able to show how the decision fits into the context in which

[5] Although a case can be made that the moral consequences of content recommendation are high

[6] The nature and possibility of moral knowledge has long been debated in philosophy. Together with his skepticism about causal knowledge, Hume likewise was skeptical about the possibility of moral knowledge [31]. However, since Socrates philosophers have also claimed that moral knowledge is not only possible but necessary for virtuous action [32]

a decision is made and how it is amenable to the moral constraints of that context.

We argue that in cases of moral knowledge, justification cannot be reliabilist. If the requirements of moral justification are defeasibility/contestability and accountability; a black-box reliabilism does not admit of either of the requirements of moral justification. The strong version of our argument is that if a decision maker is using machine learning in a high-stakes domain to aid in decisions or make decisions more efficient, the decision maker is implicitly adopting reliabilism about the machine learning output. In other words, we argue that because the black box of the model does not provide justification other than the reliabilist one, the decision maker, in virtue of using the model's output, is a reliabilist. But, reliabilism does not provide moral justification for the use of machine learning in high stakes domains. Something else is needed, namely the possibility to contest the machine learning model and to hold the humans designing and deploying the system accountable.

Core to requirements for moral justification in decision making are the notions of *defeasibility, contestability and accountability*. Contestability is at the heart of legal rights that afford individuals access to personal data and insight into the decision-making process used to classify them [30]. With current machine learning models, it is not possible to argue or reason with the model in order to "defeat" its output. A person sentenced to a longer prison term because of an algorithmic risk score cannot provide extenuating circumstances or new information to the model that might mitigate the risk score. This is because the model is opaque, its inputs are not contestable [7]. The model's output is based on aggregated statistical groups and makes decisions based on averages and statistical generalizations. Thus a machine learning risk score algorithm can explain that, for example, there might be a certain rate of re-arrest among a group of 10,000 prisoners with a very high probability; the model cannot explain why a *specific* individual is likely to be re-arrested. That is, statistical generalizations fail to capture a *causal connection* between crime and individuals [34].

Empirically, Binns [5] found that people do consider justice-related aspects of algorithmic decision-making systems, much as they do for manual decision-making processes, providing evidence for the requirement of defeasibility. Binns further found that depending on how and when they are deployed, explanations may or may not help individuals to evaluate the fairness of such decisions. In other words, we interpret this to indicate that epistemic justification may not provide the relevant moral justification. And this would follow from the fact that algorithmic decision-making systems are not defeasible in themselves, nor in the contexts of their use.

We follow the framework for contestability developed by [30] who argue that effective systems that also align with societal values require not only designs that foster in-the-moment human engagement with such systems but also governance models that support ongoing critical engagement with system processes and outputs. They define *contestability* as - the ability to challenge machine predictions. This ability is a necessary but insufficient component of moral justification.

Finally, we argue that black-box reliabilism does not admit accountability. Accountability is generally defined as the state of being responsible or answerable [10]. There is an ongoing debate about how to attribute responsibility to machine learning systems. As AI technologies gain more agency, it becomes difficult to attribute responsibility for when models fail. However, humans remain responsible for such failures because only humans still meet the criteria for moral agency and moral responsibility [6].

6 CONCLUSION

In this paper we argue that by considering machine learning as a knowledge-generating enterprise we can advance the discussion on explainability. We have introduced an epistemological argument that shifts standard calls for explainability from the epistemic inscrutability of black box models to an examination of moral justification. In other words, on the reliabilist approach to justification, standard calls for explainability focused on epistemic inscrutability of black box models are probably misplaced. As we have argued, despite the fact that what justifies the outputs of machine learning models is not accessible to us, justification is not of necessity imperilled so long as those outputs are produced by a reliable process. However, reliabilism does not provide moral justification as it does not not admit of defeasibility, contestability or accountability in situations where the human consequences of the output do not merely depend on the model's accuracy but these consequences demand a higher standard. This provides a philosophical underpinning to similar calls within the AI accountability literature to move beyond black box model explainability that only focuses on epistemic inscrutability, and instead toward models that allow for challenges to their predictions [6, 30].

To further defeasibility, contestability and accountability, in contexts where model outputs require moral as well as epistemic justification we suggest the necessity of establishing robust human processes as a moral 'wrapper' around machine outputs, processes carefully designed and tested to ensure these three criteria of moral justification are met.

Finally, a general caution, but perhaps more relevant if we adopt the reliabilist interpretation of ML models, is to be especially sensitive to population drift and to periodically check the fit of the model to the population it is intended to serve. This problem is already recognized by the machine learning field in the problem of detecting distributional drift. Even shifts in label distribution can compromise accuracy of state-of-the-art classifiers [26].

REFERENCES

[1] Kai Arulkumaran, Antoine Cully, and Julian Togelius. 2019. Alphastar: An evolutionary computation perspective. *arXiv preprint arXiv:1902.01724* (2019).

[2] Muzammil Bashir. 2019. Deep Learning Approach to Trespass Detection using Video Surveillance Data. (2019).

[3] Ruha Benjamin. 2019. *Race after technology: Abolitionist tools for the new jim code.* John Wiley & Sons.

[4] Richard Berk. 2019. *Machine learning risk assessments in criminal justice settings.* Springer.

[5] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, 377.

[6] Joanna J Bryson, Mihailis E Diamantis, and Thomas D Grant. 2017. Of, for, and by the people: the legal lacuna of synthetic persons. *Artificial Intelligence and Law* 25, 3 (2017), 273–291.

[7] Jenna Burrell. 2016. How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 2053951715622512.

[8] Richmond Campbell. 2015. Moral Epistemology. In *The Stanford Encyclopedia of Philosophy* (winter 2015 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[9] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[10] Mark Coeckelbergh. 2019. Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and engineering ethics* (2019), 1–18.

[11] Earl Conee and Richard Feldman. 1998. The generality problem for reliabilism. *Philosophical Studies* 89, 1 (1998), 1–29.

[12] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[13] Laurel Eckhouse, Kristian Lum, Cynthia Conti-Cook, and Julie Ciccolini. 2019. Layers of bias: A unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior* 46, 2 (2019), 185–209.

[14] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. 2019. A guide to deep learning in healthcare. *Nature medicine* 25, 1 (2019), 24–29.

[15] Richard Feldman. 2003. Epistemology. (2003).

[16] Edmund Gettier. 1963. Is justified true belief knowledge? *1963* (1963), 273–274.

[17] Alvin Goldman and Bob Beddor. 2016. Reliabilist Epistemology. In *The Stanford Encyclopedia of Philosophy* (winter 2016 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[18] Alvin I Goldman. 2012. *Reliabilism and contemporary epistemology: essays*. Oxford University Press.

[19] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2 (2017).

[20] Lars Hall, Petter Johansson, Betty Tärning, Sverker Sikström, and Thérèse Deutgen. 2010. Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea. *Cognition* 117, 1 (2010), 54–61.

[21] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. 2017. What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923* (2017).

[22] Gábor Horváth, Balázs Bernáth, and Gergely Molnár. 1998. Dragonflies find crude oil visually more attractive than water: multiple-choice experiments on dragonfly polarotaxis. *Naturwissenschaften* 85, 6 (1998), 292–297.

[23] Jochen Kruppa, Alexandra Schwarz, Gerhard Arminger, and Andreas Ziegler. 2013. Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications* 40, 13 (2013), 5125–5131.

[24] Jarrett Leplin. 2007. In defense of reliabilism. *Philosophical Studies* 134, 1 (2007), 31–42.

[25] David Lewis. 1996. Elusive knowledge. *Australasian journal of Philosophy* 74, 4 (1996), 549–567.

[26] Zachary C Lipton, Yu-Xiang Wang, and Alex Smola. 2018. Detecting and correcting for label shift with black box predictors. *arXiv preprint arXiv:1802.03916* (2018).

[27] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383* (2018).

[28] Spyros Makridakis. 2017. The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures* 90 (2017), 46–60.

[29] Gary Marcus. 2018. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631* (2018).

[30] Deirdre K Mulligan, Daniel Kluttz, and Nitin Kohli. 2018. Contestability and Professionals: From Explanations to Engagement with Algorithmic Systems. *Available at SSRN 3311894* (2018).

[31] Harold Noonan. 2002. *Routledge philosophy guidebook to Hume on knowledge*. Routledge.

[32] William J Prior. 2016. *Virtue and knowledge: An Introduction to ancient Greek ethics*. Routledge.

[33] Stathis Psillos. 2007. *Philosophy of science AZ*. Edinburgh University Press.

[34] Stathis Psillos. 2014. *Causation and explanation*. Routledge.

[35] Willard Van Orman Quine and Joseph Silbert Ullian. 1978. *The web of belief*. Vol. 2. Random House New York.

[36] Conrad Sachweh. 2018. General Data Protection Regulation and Explainable Machine Learning Challenges. (2018).

[37] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296* (2017).

[38] Andrew D Selbst and Solon Barocas. 2018. The intuitive appeal of explainable machines. *Fordham L. Rev.* 87 (2018), 1085.

[39] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 59–68.

[40] Stefan Thaler, Vlado Menkovski, and Milan Petkovic. 2018. Deep Learning in Information Security. *arXiv preprint arXiv:1809.04332* (2018).

[41] Jeremy Waldron. 2009. Judges as moral reasoners. *International Journal of Constitutional Law* 7, 1 (2009), 2–24.

[42] Douglas Walton. 2010. *Appeal to expert opinion: Arguments from authority*. Penn State Press.